

---

# Approximate Gaussian Integration using Expectation Propagation

---

**John P. Cunningham**

Department of Engineering  
University of Cambridge  
Cambridge, UK

**Philipp Hennig**

Department of Empirical Inference  
Max Planck Inst. f. Intelligent Systems  
Tübingen, Germany

**Simon Lacoste-Julien**

Department of Engineering  
University of Cambridge  
Cambridge, UK

## Abstract

While Gaussian probability *densities* are omnipresent in applied mathematics, Gaussian cumulative *probabilities* are hard to calculate in any but the univariate case. We offer here an empirical study of the utility of Expectation Propagation (EP) as an approximate integration method for this problem. For rectangular integration regions, the approximation is highly accurate. We also extend the derivations to the more general case of polyhedral integration regions. However, we find that in this polyhedral case, EP's answer, though often accurate, can be almost arbitrarily wrong. These unexpected results elucidate an interesting and non-obvious feature of EP not yet studied in detail, both for the problem of Gaussian probabilities and for EP more generally.

## 1 Introduction

This paper studies approximations to definite integrals of Gaussian distributions  $p_0(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{m}, K)$ . Despite the battery of convenient analytic characteristics of Gaussian *densities*, Gaussian *probabilities* are difficult to calculate: the cumulative distribution function (cdf) has no closed-form expression and is numerically challenging in high-dimensional spaces. Here we consider a generalisation of the cdf: the probability  $F(\mathcal{A}) = \text{Prob}\{\mathbf{x} \in \mathcal{A}\}$  that a draw from  $p_0(\mathbf{x})$  falls in a (possibly unbounded) *polyhedral* region  $\mathcal{A}$ . Applications of these probabilities are widespread, *e.g.* Bock and Gibbons (1996); Lesaffre and Molenberghs (1991); Thiebaut and Jacqmin-Gadda (2004); Liao et al. (2007); Gassman et al. (2002). Three machine learning problems that can be cast as Gaussian probabilities include the Bayes Point Machine (Herbrich,

2002), Gaussian Process classification (Rasmussen and Williams, 2006; Kuss and Rasmussen, 2005), and probit regression (Ochi and Prentice, 1984).

Univariate Gaussian probabilities can be calculated with machine-level precision, but no similarly powerful algorithm exists for the multivariate case. There are special cases where analytic decompositions (Plackett, 1954; Huguenin et al., 2009) or sampling methods (Lerman and Manski, 1981; Pakes and Pollard, 1989) are possible though very costly, but the most efficient general, known method is numerical integration (the recent book Genz and Bretz (2009) gives a good overview). The Genz methods represent the state of the art, but achieving high accuracy invokes substantial computational cost. Also, applications such as Bayesian model selection require *analytic* approximations of  $F(\mathcal{A})$ , usually because the goal is to optimise  $F(\mathcal{A})$  with respect to the parameters  $\{\mathbf{m}, K\}$  of the Gaussian, which requires the corresponding derivatives. These derivatives and other features are not currently offered by numerical integration methods. Thus, there remains significant work to be done to address this important problem.

Here we develop an analytic approximation to  $F(\mathcal{A})$  using Expectation Propagation (EP) (Minka, 2001a,b, 2005; Oppen and Winther, 2001, 2000) as an approximate integration method. In its most general form, this algorithm is applicable to polyhedral integration regions, which include the most commonly required hyperrectangular case (which, in turn, generalises the cdf). For hyperrectangular  $\mathcal{A}$ , the approximations are of high quality. For polyhedral  $\mathcal{A}$ , they are sometimes considerably worse. This shortcoming of EP will come as a surprise to many readers, because the differences to the rectangular case are inconspicuously straightforward. We study this interesting result, which has implications for some important applications where EP is often used. In particular, the geometric interpretation of the Gaussian probability problem provides an insightful testbed for analysing properties of EP, which has importance for approximate inference well beyond

Gaussian probabilities.

## 2 EP for Gaussian Probabilities

A prototypical aspect of Bayesian inference is that an intractable  $p(\mathbf{x})$  can be written as a product of a prior  $p_0(\mathbf{x})$  and likelihood factors  $t_i(\mathbf{x})$  such that  $p(\mathbf{x}) = p_0(\mathbf{x}) \prod_i t_i(\mathbf{x})$ . Our Gaussian probability problem is the normalisation constant of such a distribution:

$$F(\mathcal{A}) = \int_{\mathcal{A}} p_0(\mathbf{x}) d\mathbf{x} = \int p(\mathbf{x}) d\mathbf{x} = \int p_0(\mathbf{x}) \prod_{i=1}^m t_i(\mathbf{x}) d\mathbf{x} \quad (1)$$

where  $m$  can be any number of terms, greater than or less than or equal to the dimensionality  $n$  of  $p_0(\mathbf{x})$ . The factor  $t_i(\mathbf{x})$  is an indicator function defined in a particular direction  $\mathbf{c}_i$ , namely a ‘‘box function’’:

$$t_i(\mathbf{x}) = \mathbb{I}\{l_i < \mathbf{c}_i^T \mathbf{x} < u_i\} = \begin{cases} 1 & l_i < \mathbf{c}_i^T \mathbf{x} < u_i \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Note that either  $l_i$  or  $u_i$  can be infinite or inactive (not supporting the polyhedron), and thus this polyhedral definition is general. Using box functions instead of step-functions for this definition may be non-standard, but this choice will be useful in our analyses, and it reduces to hyperrectangular  $\mathcal{A}$  by setting  $m = n$  and  $\mathbf{c}_i$  to the cardinal axes.

EP approximates the intractable  $p(\mathbf{x})$  by iteratively replacing true factors  $t_i(\mathbf{x})$  with approximate factors  $\tilde{t}_i(\mathbf{x})$  from an exponential family, yielding the tractable approximate unnormalised distribution  $q(\mathbf{x}) = p_0(\mathbf{x}) \prod_i \tilde{t}_i(\mathbf{x})$ . The EP objective function is involved, but EP is *motivated* by the idea of minimising the Kullback-Leibler divergence  $D_{KL}(p||q)$ . For  $q(\mathbf{x})$  in the exponential family, this is equivalent to global moment matching (including the zeroth moment) between  $q(\mathbf{x})$  and  $p(\mathbf{x})$ .

The details of EP are by now standard in our community, so we give only a short description in the interest of brevity. At any point in the iteration, EP tracks a current approximate  $q(\mathbf{x})$  in an exponential family, and a set of *approximate factors* (or *messages*)  $\tilde{t}_i$ , also in the family. The factors are updated by constructing a *cavity* distribution

$$q^{\setminus i}(\mathbf{x}) = \frac{q(\mathbf{x})}{\tilde{t}_i(\mathbf{x})}, \quad (3)$$

and then *projecting* into the exponential family

$$\tilde{t}_i(\mathbf{x}) q^{\setminus i}(\mathbf{x}) = \text{proj}[t_i(\mathbf{x}) q^{\setminus i}(\mathbf{x})], \quad (4)$$

where this projection operation (an M-projection from information geometry (Koller and Friedman, 2009)) is

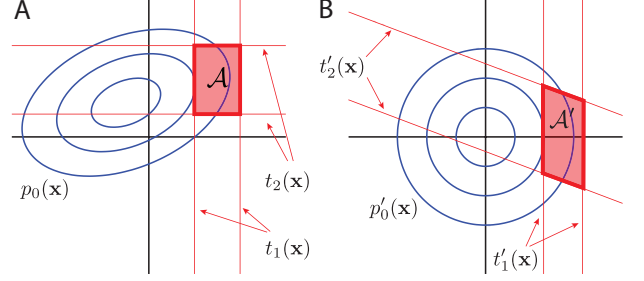


Figure 1: The effect of whitening the Gaussian probability based on the mean and covariance of the Gaussian.

defined as setting  $\tilde{t}_i(\mathbf{x})$  to the unnormalised member of the exponential family minimizing  $D_{KL}(t_i q^{\setminus i} || \tilde{t}_i q^{\setminus i})$  (Minka, 2004). This projection entails matching the sufficient statistics of  $\tilde{t}_i(\mathbf{x}) q^{\setminus i}(\mathbf{x})$  to those of  $t_i(\mathbf{x}) q^{\setminus i}(\mathbf{x})$ , which for Gaussian  $\tilde{t}_i(\mathbf{x})$  is equivalent to matching *zeroth*, *first* and *second* moments.

Our use of EP is nonstandard in two ways. First, our factors  $t_i(\mathbf{x})$  are unnormalised and thus so is  $p(\mathbf{x})$  and  $q(\mathbf{x})$  - here we use EP not to construct parametric marginal distributions, but to calculate log partition functions. Second, we note that the moment matching projection in Equation 4 is tractable for *any* rank-one factor  $t_i(\mathbf{x})$ , not simply the usual axis-aligned likelihood terms where  $\mathbf{c}_i$  are the cardinal axes. Simple rank-one updates make this algorithm computationally efficient. Critically, this generality allows us to consider probabilities  $F(\mathcal{A})$  both in axis-aligned hyperrectangular cases and arbitrary polyhedra of any orientation and any number  $m$  of constraints/faces, more or less than the dimensionality  $n$  of  $p_0(\mathbf{x})$ .

It will be useful to consider whitening the space of the problem via a substitution  $\mathbf{y} = L^{-1}(\mathbf{x} - \mathbf{m})$  where  $L$  is a Cholesky factor of the covariance  $K$ . In this transformed space, we have a standard  $\mathcal{N}(0, I)$  Gaussian, and we have a transformed set of polyhedral faces  $L^T \mathbf{c}_i$  as in Figure 1. To understand sources of error, we can then attend to properties of this transformed region. Intuition suggests that the more (less) hyperrectangular this transformed region is, the better (worse) the performance is, as hyperrectangular cases (in a white space) decompose into a product of solvable univariate problems. We then see that hyperrectangular cases are in fact not a fundamental distinction other than having the same number of constraints  $m$  as dimensions  $n$ . Considering the whitened space and transformed integration region is conceptually useful and theoretically sound, as EP with Gaussians is invariant to this linear transformation (Seeger, 2008).

The resulting algorithm, EP for multivariate Gaussian probabilities (EPMGP), yields a nice geometric interpretation. We want to integrate a Gaussian over a

polyhedron defined by several box functions, but this operation is intractable. Instead, EP allows us to replace each of those intractable truncations functions with soft Gaussian truncations  $\tilde{t}_i(\mathbf{x})$  for which multiplication is simple.

There are a number of attractive features of the EP approach which may motivate its use in many applied settings, including fast runtime, low computational overhead, and a natural ability to calculate derivatives with respect to the distribution’s parameters, tail probabilities, and first and second moments of  $p(\mathbf{x})$ .

### 3 Experiments and Results

To evaluate EP for this application, we need to establish a baseline for calculating these probabilities, and we need to choose test cases. Our baseline will be the Genz method, currently the gold standard for these calculations, which makes a series of transformations of the region  $\mathcal{A}$  and the Gaussian  $p_0(\mathbf{x})$  to enable accurate numerical integration. The integrand of Equation 1 is transformed to the unit cube, and heuristic choices on integration order are made before quasi-random integration or lattice-point rules are used (Genz and Bretz, 2009). For our comparisons to EPMGP, we used the Genz method with  $5 \times 10^5$  lattice points.

We note that Monte Carlo methods are not particularly well suited to this problem: rejection samplers from  $p_0(\mathbf{x})$ , importance samplers over the polyhedron  $\mathcal{A}$  (“pinball” or “hit-and-run” MCMC algorithms (Herbrich, 2002)), and elliptical slice sampling (Murray et al., 2010) do not perform favorably (the latter is best among samplers, but still well beneath EP and Genz). In light of the accuracy of the Genz algorithms, we will not consider samplers further here.

We next describe choosing our problem cases. To define a Gaussian  $p_0(\mathbf{x})$ , we draw eigenvalues from an exponential distribution and eigenvectors uniformly from the unit hypersphere to form the covariance  $K$ . This procedure produces a more interesting range of  $K$  - in particular a better spread of condition numbers - than using a Wishart or similar. We set the mean  $\mathbf{m} = \mathbf{0}$  without loss of generality.

In the hyperrectangular integration cases (where dimension  $n$  equals number of constraints  $m$ ),  $\mathcal{A}$  can be defined by the upper and lower bounds  $\mathbf{u}$  and  $\mathbf{l}$ , which we defined by taking a draw from  $p_0(\mathbf{x})$  and adding and subtracting uniform random lengths that scale with dimensionality  $n$ . Having the region size scale with the dimensionality  $n$  helps to prevent the probabilities  $F(\mathcal{A})$  from becoming vanishingly small with increasing dimension (which is handled fine by EP but problem-

atic for the Genz method). In the arbitrary polyhedral case, the same procedure is repeated for drawing  $\mathbf{u}$  and  $\mathbf{l}$  for all  $m$  constraints. We also choose axes of orientation, which are unit-norm vectors  $\mathbf{c}_i \in \mathbb{R}^n$  for each of the  $m$  factors, again uniformly from the unit hypersphere.

It is also useful to consider special cases where we can analytically find the true probability  $F(\mathcal{A})$ . Orthants generalise quadrants to high dimension, for example  $\{\mathbf{x} \in \mathbb{R}^n : x_i > 0 \forall i\}$ . For zero-mean Gaussians in  $\mathbb{R}^2$ , calculating the probability  $F(\mathcal{A})$  when  $\mathcal{A}$  is an orthant is a simple geometry problem, and there are also geometric solutions in  $\mathbb{R}^3$ ,  $\mathbb{R}^4$ , and  $\mathbb{R}^5$  (Genz and Bretz, 2009; Sinn and Keller, 2010).

#### 3.1 EP results for hyperrectangular and polyhedral integration regions

Figure 2A-C shows the empirical results for hyperrectangular integration regions. For each dimension  $n = \{2, 3, 4, 5, 10, 20, 50, 100\}$ , we chose 250 random Gaussians  $p_0(\mathbf{x})$  and random regions  $\mathcal{A}$ , and we calculated the relative error between the EPMGP and Genz estimates. Each of these errors is plotted against dimensionality  $n$  as a point in Figure 2A, with added horizontal jitter to show the distribution of points. The blue line and error bars represent the median and  $\{25\%, 75\%\}$  quartiles. The black line below the EP results is the median error estimate given by the Genz method. Since this line is almost always a few orders of magnitude below the distribution of EP errors, it is safe to use the high accuracy Genz method as a proxy to ground truth. Figure 2A demonstrates that the EP method gives a reliable estimate of Gaussian probabilities with hyperrectangular integration regions, with relative errors typically on the order of  $10^{-4}$  and with individual errors rarely in excess of 1%.

Figure 2B plots the errors by the value of  $\log Z$  itself, where we use the  $\log Z$  calculated from Genz. On this panel we can also plot a 1% error for  $Z \approx F(\mathcal{A})$ , as the log scale can suggest misleadingly large or small errors. Though it is a common intuition that this method should have high error in tail probabilities, there is instead a weakly positive or nonexistent correlation of error with  $\log Z$ . Panel C repeats the same errors but plotted by the condition number of the covariance  $K$  and shows pronounced trend, which is expected by considering Figure 1.

Figure 2D-F has the same setup as 2A-C, except we test over polyhedral integration regions, which amounts to additionally drawing  $m = n$  random polyhedral faces  $\mathbf{c}_i$ . The data tell a similar story but with surprisingly higher errors. The black line in Panel D shows the median EP errors in the axis aligned case

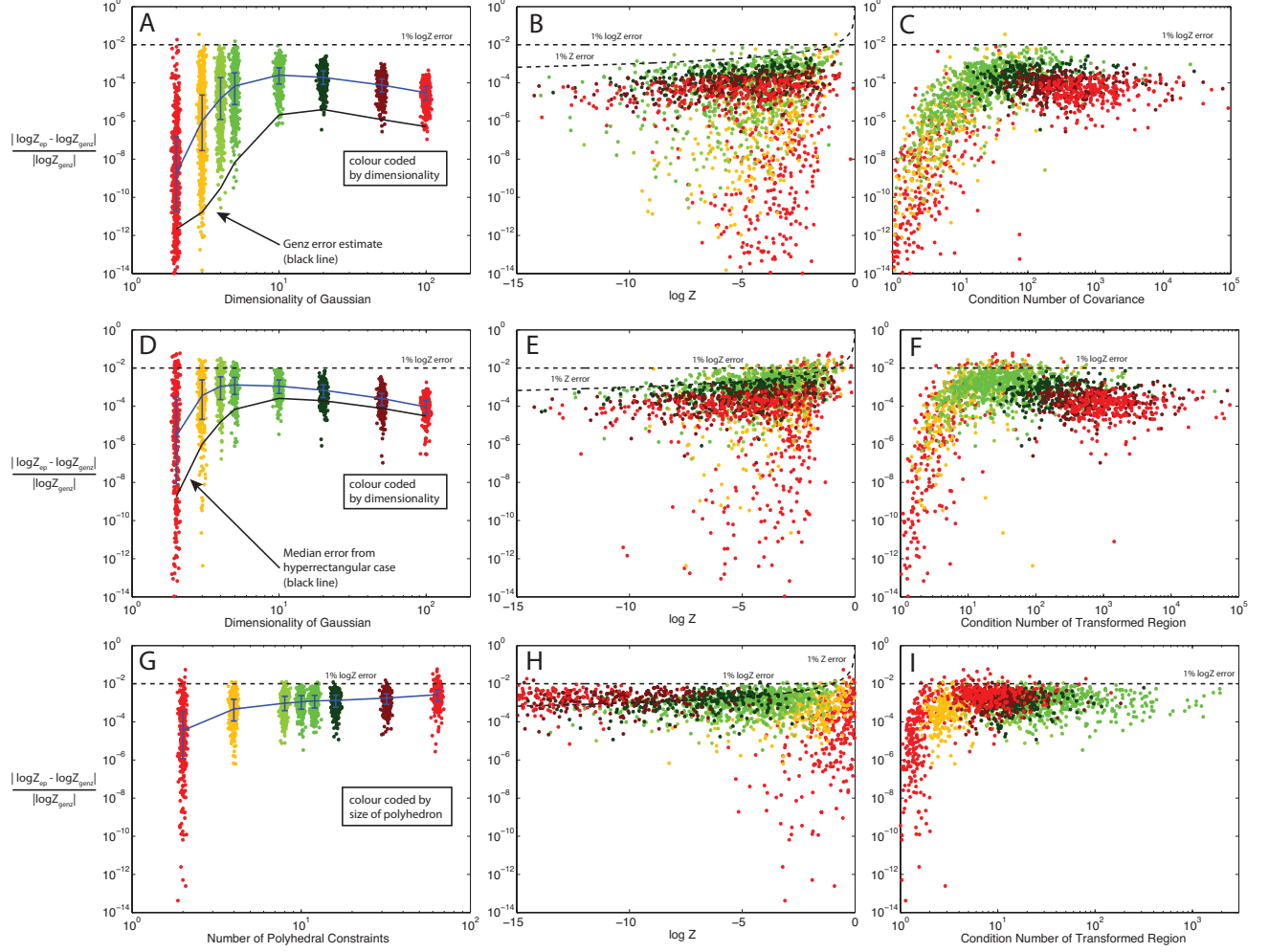


Figure 2: Empirical results for EPMGP with hyperrectangular and polyhedral integration regions. See text.

(*i.e.*, the blue line in Figure 2A). The purpose of this line is to show that indeed for the same Gaussian cases, the error rate of polyhedral EPMGP is typically an order of magnitude or two higher than in the hyperrectangular case. Genz errors are not shown. Again Figure 2E shows a modest positive effect in  $\log Z$ . The only other change is Figure 2F, where we use the concept of whitening the space to consider the transformed region integrated over a standard Gaussian as in Figure 1. We consider this region as a matrix  $C'$  with columns equal to  $L^T \mathbf{c}_i$ . To test the transformed region's proximity to  $I$  (a decomposable problem with no EP error), we use the condition number of  $C'$ . Conveniently, in the hyperrectangular case this metric is the square root of the condition number of the covariance  $K$ . This allows us to compare Figures 2C and 2F, since this square root is just a constant scaling on the log scale shown. Figure 2F confirms our intuition that less hyperrectangular transformed integration regions will have higher

error<sup>1</sup>.

The sensitivity of error to the number of polyhedral constraints  $m$  is also an important question. Figure 2G-I gives these empirical results: Panel G plots errors by number of polyhedral face constraints  $m$  instead of by Gaussian dimension  $n$ . In this row of the figure, all cases use  $n = 10$  dimensions, and we show polyhedral sizes of  $m = \{2, 4, 8, 10, 12, 16, 32, 64\}$ . Larger numbers of constraints/polyhedral faces do imply larger error, trending up by roughly an order of magnitude or two. The errors shown in Figure 2H seem largely invariant to the value  $\log Z$  itself. However, Panel I still indicates some upward trend, as we would expect.

Figure 2D-I, namely the arbitrary polyhedral cases, show error rates that may in many cases be unaccept-

<sup>1</sup>There are a number of other sensible metrics on the transformed region, and we have produced similar plots with Frobenius and  $l_1$  norms of the scaled Gram matrix  $\frac{1}{n} C'^T C'$ . The error trend is the same.

able. Though EP still performs well often, the reliability of the algorithm is not what it is in the hyper-rectangular case of Figure 2A-C.

Next, we calculate true orthant probabilities analytically: Figure 3 shows errors for both the EP method (colour) and the Genz method (grayscale). The two panels are the usual errors plotted as a function of condition number of covariance (Panel A) and the true probability  $\log Z$  (Panel B). The four cases shown are orthant probabilities at  $n = \{2, 3, 4, 5\}$ . First, we note that there is a clear separation between the Genz errors and the EP errors, which helps to solidify the earlier claim that the Genz numerical answer can be used as a reasonable proxy to ground truth.

Second, it is interesting to note that there is significant structure in the EP errors with orthant probabilities when plotted by  $\log Z$  in Panel B. Each dimension has a “V” shape of errors. This can be readily explained by reviewing what an orthant probability is. For a zero-mean Gaussian, an orthant probability in  $\mathbf{R}^2$  is simply the mass of one of the quadrants. If that quadrant has a mass of  $\log Z = \log 0.25$ , then the correlation must be white, and hence EP will produce a highly accurate answer (the problem decomposes). Moving away from white correlation, EP will produce error. This describes the shape of the red curve for  $n = 2$ , which indeed is minimised at  $\log 0.25 \approx -1.39$ . The same argument can be extended for why there should be a minimum in  $\mathbf{R}^3$  at  $\log Z = \log 0.125$ , and so on.

In summary, the empirical results of Figures 2 and 3 indicate that EPMGP is a successful candidate algorithm for Gaussian probabilities. The error rate is non-zero but generally quite low, with median errors less  $10^{-4}$  and individual errors rarely in excess of 1% across a range of dimensions, which may be acceptable in many applied settings.

### 3.2 Pathological cases

Figure 4 shows pathological cases designed to illustrate the shortcomings of EP. In all cases, we want to integrate the  $\mathcal{N}(0, I)$  Gaussian in  $n = 2$  dimensions over the  $[-1, 1] \times [-1, 1]$  box, which is simply the product of two axis-aligned univariate box functions ( $t_1(x_1) = \mathbb{I}\{-1 < x_1 < 1\}$  and  $t_2(x_2) = \mathbb{I}\{-1 < x_2 < 1\}$ ). The  $y$ -axis shows the errors as usual, and the  $x$ -axis is a feature chosen to create large errors.

First, when looking at a transformed integration region, EP error may derive from approximate factors  $\tilde{t}_i(\mathbf{x})$  being not orthogonal, and thus overcounting the mass when forming the EP approximation  $q(\mathbf{x})$ . An extreme, contrived example of this is adding two repeats of identical factors  $t_i(\mathbf{x})$ . Though the integration region is unchanged, EP is not invariant to this

change, as it still must make an approximate factor for each true factor. Figure 4A shows this **redundancy** pathology. We use EP to solve the same probability with increasing numbers of redundant factors, adding up to 1000 copies of the same integration region. Mathematically, we go from using EP to solve  $\int p_0(\mathbf{x})t(x_1)t(x_2)d\mathbf{x}$ , to using EP to solve  $\int p_0(\mathbf{x})t(x_1)^{1000}t(x_2)^{1000}d\mathbf{x}$ . Though the true  $F(\mathcal{A})$  is unchanged, EP becomes arbitrarily bad. EP in this redundancy case is *underestimating* the true probability, for reasons explained in the next section.

Second, EP error may derive from the approximate factors accounting for mass inappropriately. When two  $t_i(\mathbf{x})$  are not orthogonal, moment matching will include mass that is outside the integration region. Figure 4B shows an integration region that is still the  $[-1, 1] \times [-1, 1]$  box, but that can be described by the intersection of two boxes of any size, as the cartoon shows. However, EP must consider the mass that exists in each true box factor *individually*. Hence we expect and indeed see that EP will not be invariant to this change, and further that EP should overestimate the true probability when there is **extra mass**, which is the term we use to describe this pathology of EP. Note that the case where there is no extra mass - Panel B at left - is still a case where there are two redundant boxes, and hence we expect the underestimation error from Panel A (corresponding points circled in blue). Lastly, Figure 4C shows that preprocessing the constraints (by removing redundant factors or tightening factor limits) can not fix these pathologies: increasing numbers of slightly rotated boxes minimally describe  $\mathcal{A}$ , but the same magnitude of errors is seen.

These redundancy and extra mass issues are of course two effects of a single underlying problem: EP enforces a Gaussian approximation to a hard box-function factor, which is a highly non-Gaussian form (but better than a conventional halfspace step function). This effect is particularly pronounced for sharp box factors, but not limited to them: factors with weaker or heavier tails, and non-symmetric factors, all have similar issues. Thus redundancy and extra mass generally go hand in hand: if there are likelihood terms that are not orthogonal with respect to the prior covariance  $K$ , they will each have to consider both mass already considered by another factor (redundancy) and mass that lies outside the actual polyhedron (extra mass).

## 4 Correcting EP Errors

These empirical results suggest two ways to improve EP for multivariate Gaussian probabilities, and they shed further light on the underpinnings of EP’s failure modes.

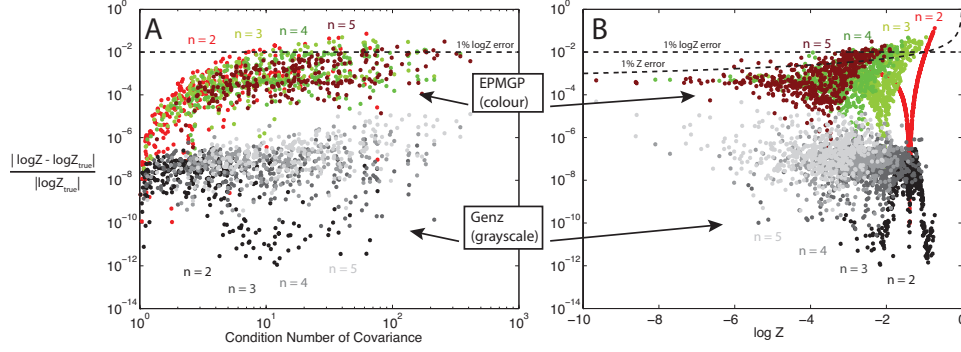


Figure 3: Empirical results for orthant probabilities. See text for description.

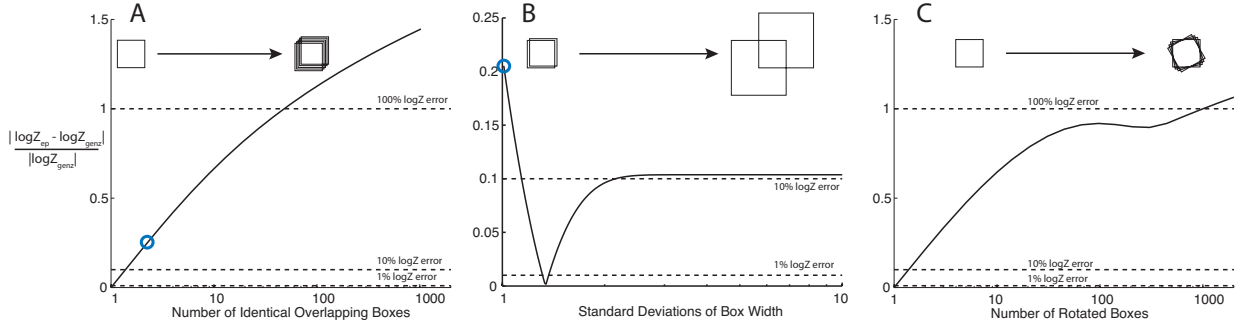


Figure 4: Empirical results from intentionally pathological cases. See text for description.

#### 4.1 $\alpha$ -divergence perspective

The redundancy and extra mass problems can be viewed as changing the effective  $\alpha$ -divergence that EP minimises. The  $\alpha$ -divergence  $D_\alpha(p \parallel q)$  is a generalisation of the KL-divergence, and its relationship to EP and other variational schemes has been studied (Minka, 2005). The Power-EP algorithm is an extension of EP where the factors are raised to a specific power  $\alpha_i$  before moment matching (Minka, 2004). One of the motivations for this algorithm is that raising the factor  $t_i$  to a power makes the projection more tractable. That paper also notes that running EP on a model where each factor  $t_i$  is repeated  $n_i$  times is equivalent to running Power-EP (with the factors unrepeated) with  $\alpha_i = 1/n_i$ . Importantly, the Power-EP algorithm can also be derived as the minimisation of local  $\alpha_i$ -divergences. This repetition of factors is precisely the “redundancy” issue that we created by repeating constraints  $n_i$  times in Figure 4A. Our EP algorithm in the pure redundant case can thus be interpreted as running Power-EP with  $\alpha_i = 1/n_i$ .

First, we note that the zeroth moment  $Z$  is always *underestimated* by  $\alpha$ -divergence with  $\alpha < 1$ . Also,  $Z$  is only correctly estimated for  $\alpha = 1$  (KL-divergence), and it is *overestimated* for *inclusive*  $\alpha$ -divergences which have  $\alpha > 1$ . This provides a theoretical ex-

planation for the systematic underestimation of  $Z$  in the redundancy case of Figure 4A: EP in this case is the same as Power-EP with  $\alpha_i = 1/n_i < 1$ . As in regular EP, a subtle aspect here is that Power-EP minimises the  $\alpha$ -divergence *locally*, whereas the overestimation/underestimation of the zeroth moment  $Z$  holds for *global*  $\alpha$ -divergence minimisation, and the relationship between the two is not yet fully understood. However, our construction in Figure 4A was a fully factorised case, so there is direct correspondence between global and local divergences, and as such we did systematically observe underestimation in these experiments. Our experiments indicate similar underestimation even when there is not direct global-local correspondence.

Second, this relationship between EP and  $\alpha$ -divergence gives us a notion of *effective*  $\alpha$ -divergence  $\alpha_{\text{eff}}$ . In particular, running EP with the factor  $t_i$  repeated  $m_i$  times is the same as local  $\alpha$ -divergence minimisation by Power-EP with  $\alpha_{\text{eff}} = 1/m_i$ . When the constraints are almost-repeated (they have a large dot product, but are not fully colinear, such as Figure 4C), we lose the exact correspondence between EP and Power-EP, but we could still imagine that EP corresponds to Power-EP with a suitable  $\frac{1}{m_i} < \alpha_{\text{eff}} < 1$  in this case also (using a continuity argument).



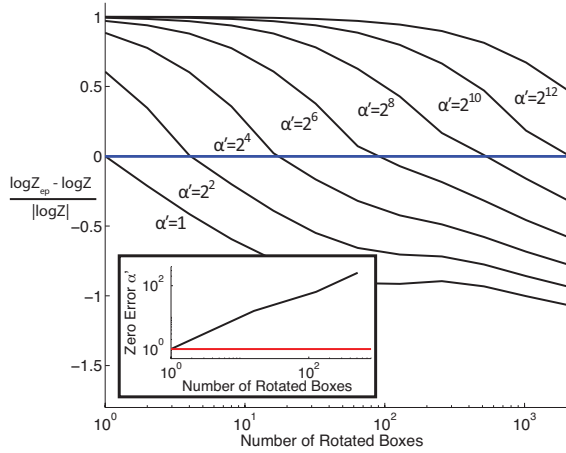


Figure 5:  $\alpha'$  corrections to the rotated box example. The line  $\alpha' = 1$  (no correction) is the same curve as Figure 4C.

With these facts, we can try to correct for this  $\alpha_{\text{eff}}$  by using Power-EP with  $\alpha' = 1/\alpha_{\text{eff}}$  instead of standard EP. Thus we can use the above arguments directly in Figure 4A to remove the errors entirely. For box factors that are repeated  $m_i$  times, we correct with  $\alpha'_i = 1/\alpha_{\text{eff}} = m_i$ . Empirical findings confirm this theoretical treatment, and the errors in Figure 4A are entirely removed by this Power-EP correction. In Figure 5 we repeat the experiment of Figure 4C, but for a range of  $\alpha'$  corrections. We use a single correction for all factors  $t_i(\mathbf{x})$ , though in general each factor may be corrected individually. The  $\alpha' = 1$  curve in Figure 5 is the same as Figure 4C (up to a sign change). We then plot the same curve for many different corrections, and those black curves indicate that indeed there is an optimal  $\alpha'$  correction that allows the EP estimate to be error free. In the inset panel we plot those optimal  $\alpha'$  values, which shows, for a given number of rotated boxes ( $x$ -axis, log scale), the corrective  $\alpha'$  required ( $y$  axis, log scale) to remove EP error entirely from the result. Here the curve is not linear, nor are the optimal  $\alpha' = m$  (though reasonably close, since slightly rotated boxes are nearly repeats). These results are not surprising, since there should be some error-free value in many cases. Nonetheless we see that even choosing a sensible  $\alpha'$  correction can considerably improve the EP result in this redundancy example.

Next, the extra mass issue can also be viewed in terms of  $\alpha$ -divergences, though the connection is not as rigorous as the redundancy issue. The extra mass problem is one of inclusivity: the Gaussian approximate factor is including mass that it should not (as that mass lies outside the true polyhedron), which is a property of *inclusive*  $\alpha$ -divergences as mentioned previously, *i.e.*, divergences with  $\alpha > 1$ . These we would expect to

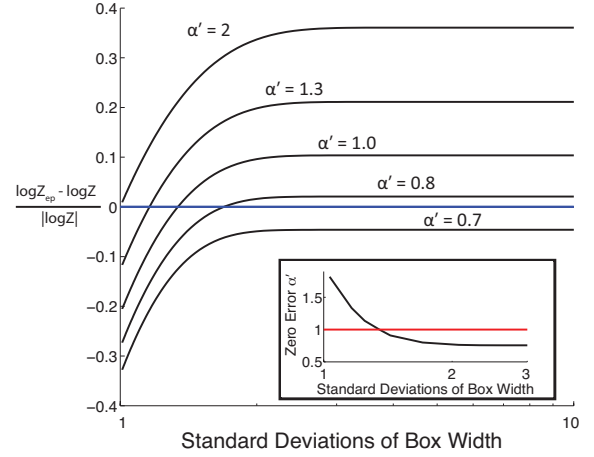


Figure 6:  $\alpha'$  corrections to the extra mass example. The line  $\alpha' = 1$  (no correction) is the same curve as Figure 4B.

overestimate the true probability, and indeed this is what we see in Figure 4B. We can also consider correcting for these  $\alpha_{\text{eff}}$  with Power EP. However, the situation is slightly more complicated here as the extra mass problem involves understanding how  $\alpha_{\text{eff}}$  changes with the decay of the Gaussian distribution. Results showing this are shown in Figure 6. This figure tells the same story as Figure 5: a sensible choice of correction term will remove much of the EP error.

Of course, the challenge is to come up with a theory or principled heuristics for choosing  $\alpha'_i$  in general polyhedral problems, as intuition tells us there should be no straightforward answer here. For almost-repeated constraints and polyhedra in general, we have attempted to similarly improve EP by calculating a suitable  $\alpha_{\text{eff}}$  term from problem parameters (the region and the Gaussian), and using Power-EP. Our early investigation of this correction has given some promising results, often reducing errors significantly. However, because often  $\alpha_{\text{eff}} < 1$  and hence the correction term is greater than unity (unlike standard Power-EP), this EP algorithm encounters frequent convergence and stability issues, even with significant damping on the outbound messages (which does *not* alter the solution or  $\alpha_{\text{eff}}$ ). This obstacle suggests two areas of future work: first, better characterisation of  $\alpha_{\text{eff}}$  from geometric quantities such as dot products of constraints. Second, addressing convergence issues via a provably convergent double-loop approach to EP (Oppor and Winther, 2005; Seeger and Nickish, 2011).

To summarise, both the redundancy and extra mass problems can be cast in the language of effective  $\alpha$ -divergence. We know that minimising  $D_\alpha$  only returns moment matching when  $\alpha = 1$  (KL-divergence), and thus any efforts to drive  $\alpha_{\text{eff}}$  closer to 1 should improve

all the results seen here.

## 4.2 Perturbation perspective

Another important direction of work in correcting EP is approximations to the remainder term  $R = \frac{\log Z}{\log Z_{EP}}$  (Paquet et al., 2009; Oppen et al., 2009; Cseke and Heskes, 2011). By making various expansions to this remainder term, we might hope to generally reduce EP errors. A particularly relevant expansion is from Paquet et al. (2009) and uses the error terms  $\epsilon_i(\mathbf{x}) = \frac{t_i(\mathbf{x})}{Z_i t_i(\mathbf{x})} - 1$ , such that the correction term  $R$  becomes a finite series of moments over products of increasing numbers of true factors and approximating factors. Of course this is intractable in general. However, in our problem setting these moments are again multivariate probabilities of increasing order, which allows us to consider a variety of interesting schemes.

Our current direction of work is to use the Genz method in low-dimensions as a refinement to EP. This effectively allows us to capture all the benefits of the EP method, which scales well with dimension and  $\log Z$ , while using numerical quadrature where it is fast and extremely accurate, namely in the low dimensional case. This creates a hybrid EP-Genz method, where we can trade between the two approaches on a spectrum determined by the depth of the expansion. An important caveat of these expansion methods is that they rely on an already-good approximation from EP. However, our results from Figure 2 indicate that this is likely acceptable in most all cases, as EP is already highly accurate. At this time we do not have preliminary results to report on this approach.

## 5 Discussion

We have presented an EP algorithm constructing analytic approximations to polyhedral integrals on multivariate Gaussian distributions. Such approximations highlight interesting connections between approximate inference and approximate integration, and these approximations are of value for a number of applications in machine learning and throughout science and applied statistics. We show that under a variety of circumstances, relative EP errors rarely exceed 1%, and median relative errors are typically two to four orders of magnitude smaller. While we spent much of the results exploring the regions of poor performance and possible corrections, the general strength of the method should not be neglected. EPMGP also has some nice features such as analytical derivatives, fast runtime, and a natural ability to compute tail probabilities. On the other hand, existing numerical meth-

ods, in particular the Genz method, typically demonstrate high numerical accuracy.

Our results put a qualifier on the results of previous work using EP for evidence estimation, such as the Bayes Point Machine (Herbrich, 2002; Minka, 2001a,b), which was thoroughly investigated only for a relatively low dimensional case ( $m = 3$ ,  $n = 2$  in our notation) and for prediction accuracy, not ground truth. Our results caution that in such applications, EP should be used with care in problems where constraints outnumber dimensions, particularly  $m \gg n$ .

Another important existing connection point is Gaussian Process classification (Rasmussen and Williams, 2006; Kuss and Rasmussen, 2005). When a probit likelihood (namely a univariate Gaussian cdf) is used, the data likelihood becomes a Gaussian probability problem. The same is true of probit regression (Ochi and Prentice, 1984), where changing the representation of the more standard prior and likelihood reveals a high-dimensional Gaussian probability. Though these problems are usually not considered Gaussian probabilities, it is another point of broad importance for EP and for Gaussian probabilities. Furthermore, as EP is often used in a different way in these problems, exploring the similarities between EP for Gaussian probabilities and EP for problems that can be cast as Gaussian probabilities is an interesting subject for future investigation.

Here we have investigated properties for EP as applied to Gaussian integration, but understanding how much of these issues stems from EP itself and how much from the application is an important clarification. Gaussian probabilities are particularly instructive, as they encourage the geometric interpretation that we have maintained throughout this paper. However, the issues of redundancy and extra mass are not specific to strict box functions, and the interpretation in terms of  $\alpha_{\text{eff}}$  is no less applicable in the case of smooth functions.

Perhaps the most interesting aspect of this work is that the Gaussian probability problem, by its simple geometric interpretation, allows clear investigation into the strengths and weaknesses of EP, which bears relevance to applications well beyond just Gaussian probabilities. As EP becomes a more popular approximate inference method, this cautionary result suggests that EP should not be trusted blindly in new applications.

## Acknowledgements

We thank Tom Minka, Thore Graepel, Ralf Herbrich, Carl Rasmussen, Zoubin Ghahramani, Ed Snelson, Ulrich Paquet, and David Knowles for helpful discussions. JPC was supported by the UK Engineer-



ing and Physical Sciences Research Council (EPSRC EP/H019472/1); PH and SLJ were supported by a grant from Microsoft Research Ltd.

## References

- Bock, R. D. and Gibbons, R. D. (1996). High-dimensional multivariate probit analysis. *Biometrics*, 52(4):1183–1194.
- Cseke, B. and Heskes, T. (2011). Approximate marginals in latent Gaussian models. *Journal of Machine Learning Res.*, 12:417–454.
- Gassman, H. I., Deak, I., and Szantai, T. (2002). Computing multivariate normal probabilities: A new look. *J Comp Graph Stats*, 11(4):920–949.
- Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and t probabilities*. Springer.
- Herbrich, R. (2002). *Learning Kernel Classifiers*. MIT Press.
- Huguenin, J., Pelgrin, F., and Holly, A. (2009). Estimation of multivariate probit models by exact maximum likelihood. Technical report, University of Lausanne, Institute of Health and Economics Management (IEMS).
- Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.
- Kuss, M. and Rasmussen, C. (2005). Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Res.*, 6:1679–1704.
- Lerman, S. and Manski, C. (1981). On the use of simulated frequencies to approximate choice probabilities. *Structural analysis of discrete data with econometric applications*, 10:305–319.
- Lesaffre, E. and Molenberghs, G. (1991). Multivariate probit analysis: a neglected procedure in medical statistics. *Statistics in Medicine*, 10(9):1391–1403.
- Liao, X., Li, H., and Carin, L. (2007). Quadratically gated mixture of experts for incomplete data classification. In *Proceedings of the 24th International Conference on Machine Learning*.
- Minka, T. P. (2001a). Expectation Propagation for approximate Bayesian inference. *Uncertainty in AI*, pages 362–369.
- Minka, T. P. (2001b). *A family of algorithms for approximate Bayesian inference*. MIT Ph.D. Thesis.
- Minka, T. P. (2004). Power EP. Technical report, Microsoft Research.
- Minka, T. P. (2005). Divergence measures and message passing. *Microsoft Research Tech. Report MSR-TR-2005-173*.
- Murray, I., Adams, R. P., and MacKay, D. J. C. (2010). Elliptical slice sampling. *Proc. 13th Intl Conference on AISTATS*.
- Ochi, Y. and Prentice, R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika*, 71(3):531–543.
- Opper, M., Paquet, U., and Winther, O. (2009). Improving on expectation propagation. In *Advances in Neural Information Processing Systems*. MIT Press.
- Opper, M. and Winther, O. (2000). Gaussian processes for classification: Mean field algorithms. *Neural Comp.*, 12:2655–2684.
- Opper, M. and Winther, O. (2001). From naive mean field theory to the TAP equations. In Opper, M. and Saad, D., editors, *Advanced Mean Field Methods: Theory and Practice*. MIT Press, Cambridge, MA.
- Opper, M. and Winther, O. (2005). Expectation consistent approximate inference. *Journal of Machine Learning Res.*, 6:2177–2204.
- Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, 57(5):1027–1057.
- Paquet, U., Winther, O., and Opper, M. (2009). Perturbation corrections in approximate inference: Mixture modeling applications. *Journal of Machine Learning Res.*, 10:1263–1304.
- Plackett, R. L. (1954). A reduction formula for normal multivariate integrals. *Biometrika*, 41(3–4):351–360.
- Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge.
- Seeger, M. (2008). Expectation propagation for exponential families. *Technical Report, U.C. Berkeley*.
- Seeger, M. and Nickish, H. (2011). Fast convergent algorithms for Expectation Propagation approximate Bayesian inference. *Proc. 14th Intl Conference on AISTATS*.
- Sinn, M. and Keller, K. (2010). Covariances of zero crossings in Gaussian processes. *Theory Prob and Appl.* To appear.
- Thiebaud, R. and Jacqmin-Gadda, H. (2004). Mixed models for longitudinal left-censored repeated measures. *Comput Methods Programs Biomed*, 74(3):255–260.

## Appendix: details for algorithm and equations

For completeness, we include here a more detailed description of the EPMGP algorithm:

---

**Algorithm 1** EPMGP: Calculate  $F(\mathcal{A})$ , the probability of  $p_0(\mathbf{x})$  on the region  $\mathcal{A}$ .

---

- 1: Initialise  $q(\mathbf{x})$  ( $\{Z, \boldsymbol{\mu}, \Sigma\}$ ; typically  $q(\mathbf{x}) = p_0(\mathbf{x})$ ) and factors  $\tilde{t}_i(\mathbf{x})$  with zero precision.
  - 2: **while**  $q(\mathbf{x})$  has not converged **do**
  - 3:   **for**  $i \leftarrow 1 : m$  **do**
  - 4:     form cavity local  $q^{\setminus i}(\mathbf{x})$  by Equation 5.
  - 5:     calculate moments of  $q^{\setminus i}(\mathbf{x})t_i(\mathbf{x})$  by Equations 6-7.
  - 6:     choose  $\tilde{t}_i(\mathbf{x})$  so  $q^{\setminus i}(\mathbf{x})\tilde{t}_i(\mathbf{x})$  matches above moments by Equation 8.
  - 7:   **end for**
  - 8:   update  $\boldsymbol{\mu}, \Sigma$  with new  $\tilde{t}_i(\mathbf{x})$  by Equation 9.
  - 9: **end while**
  - 10: calculate  $Z$ , the total mass of  $q(\mathbf{x})$  using Equation 10.
  - 11: **return**  $Z$ , the approximation of  $F(\mathcal{A})$ .
- 

Given a current EP approximation  $q(\mathbf{x}) = p_0(\mathbf{x}) \prod_i \tilde{t}_i(\mathbf{x}) = Z\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$  and individual approximate factors  $\tilde{t}_i(\mathbf{x}) = \tilde{Z}_i \mathcal{N}(\mathbf{c}_i^T \mathbf{x}; \tilde{\mu}_i, \tilde{\sigma}_i^2)$ , we have the following rule for calculating the EP cavity distribution of Equation 3, (which is actually only needed along the  $\mathbf{c}_i$  direction in our rank-one case:  $q^{\setminus i}(\mathbf{c}_i^T \mathbf{x}) = Z_{\setminus i} \mathcal{N}(\mathbf{c}_i^T \mathbf{x}; \mu_{\setminus i}, \sigma_{\setminus i}^2)$ ) – it is a Gaussian with parameters:

$$\mu_{\setminus i} = \sigma_{\setminus i}^2 \left( \frac{\mathbf{c}_i^T \boldsymbol{\mu}}{\mathbf{c}_i^T \Sigma \mathbf{c}_i} - \frac{\tilde{\mu}_i}{\tilde{\sigma}_i^2} \right), \quad \text{and} \quad \sigma_{\setminus i}^2 = ((\mathbf{c}_i^T \Sigma \mathbf{c}_i)^{-1} - \tilde{\sigma}_i^{-2})^{-1}. \quad (5)$$

Setting the  $\mathbf{c}_i$  to the cardinal axes returns a hyperrectangular (axis-aligned) integration region and a more conventional EP algorithm with the familiar cavity parameter forms (*e.g.*, Equation 3.56 in Rasmussen and Williams (2006)).

For the projection step given in Equation 4, we first need to compute the zeroth, first and second moments  $\{\hat{Z}_i, \hat{\mu}_i \mathbf{c}_i, \hat{\sigma}_i^2 \mathbf{c}_i \mathbf{c}_i^T\}$  of  $q^{\setminus i}(\mathbf{x})t_i(\mathbf{x})$  which can be done exactly using the error function:

$$\hat{Z}_i = \frac{1}{2} \left( \text{erf}(\beta) - \text{erf}(\alpha) \right), \quad \hat{\mu}_i = \mu_{\setminus i} + \frac{1}{\hat{Z}_i} \frac{\sigma_{\setminus i}}{\sqrt{2\pi}} \left( \exp\{-\alpha^2\} - \exp\{-\beta^2\} \right), \quad (6)$$

$$\hat{\sigma}_i^2 = \mu_{\setminus i}^2 + \sigma_{\setminus i}^2 + \frac{1}{\hat{Z}_i} \frac{\sigma_{\setminus i}}{\sqrt{2\pi}} \left( (l_i + \mu_{\setminus i}) \exp\{-\alpha^2\} - (u_i + \mu_{\setminus i}) \exp\{-\beta^2\} \right) - \hat{\mu}_i^2, \quad (7)$$

where we have everywhere above used the shorthand  $\alpha = \frac{l_i - \mu_{\setminus i}}{\sqrt{2\sigma_{\setminus i}^2}}$  and  $\beta = \frac{u_i - \mu_{\setminus i}}{\sqrt{2\sigma_{\setminus i}^2}}$ . These moments are used to update  $\tilde{t}_i(\mathbf{x})$  in the usual way, using standard manipulation of Gaussian random variables:

$$\begin{aligned} \tilde{t}_i(\mathbf{x}) &= \tilde{Z}_i \mathcal{N}(\mathbf{c}_i^T \mathbf{x}; \tilde{\mu}_i, \tilde{\sigma}_i^2), \\ \text{where } \tilde{\mu}_i &= \tilde{\sigma}_i^2 (\hat{\sigma}_i^{-2} \hat{\mu}_i - \sigma_{\setminus i}^{-2} \mu_{\setminus i}), \quad \tilde{\sigma}_i^2 = (\hat{\sigma}_i^{-2} - \sigma_{\setminus i}^{-2})^{-1}, \\ \tilde{Z}_i &= \hat{Z}_i \sqrt{2\pi} \sqrt{\sigma_{\setminus i}^2 + \tilde{\sigma}_i^2} \exp\left\{ \frac{1}{2} (\mu_{\setminus i} - \tilde{\mu}_i)^2 / (\sigma_{\setminus i}^2 + \tilde{\sigma}_i^2) \right\}. \end{aligned} \quad (8)$$

Now, by the definition of the approximation, we can calculate the new approximation  $q(\mathbf{x})$  as the product  $q(\mathbf{x}) = p_0(\mathbf{x}) \prod_i \tilde{t}_i(\mathbf{x})$ :

$$q(\mathbf{x}) = Z\mathcal{N}(\boldsymbol{\mu}, \Sigma), \quad \text{where } \boldsymbol{\mu} = \Sigma \left( K^{-1} \mathbf{m} + \sum_{i=1}^m \frac{\tilde{\mu}_i}{\tilde{\sigma}_i^2} \mathbf{c}_i \right), \quad \Sigma = \left( K^{-1} + \sum_{i=1}^m \frac{1}{\tilde{\sigma}_i^2} \mathbf{c}_i \mathbf{c}_i^T \right)^{-1}, \quad (9)$$

where  $\mathbf{m}, K$  are the parameters of  $p_0(\mathbf{x})$ . These forms can actually be efficiently and numerically stably calculated in quadratic computation time in our implementation using the natural parameterisation of the Gaussian rather than the mean parameters, given a Cholesky decomposition of  $K$ . Lastly, once the algorithm has converged,

we can calculate the normalisation constant of  $q(\mathbf{x})$  which is our quantity of interest - the approximation of the probability  $F(\mathcal{A})$ :

$$\begin{aligned} \log Z &= -\frac{1}{2} (\mathbf{m}^T K^{-1} \mathbf{m} + \log|K|) \\ &\quad + \sum_{i=1}^m \left( \log \tilde{Z}_i - \frac{1}{2} \left( \frac{\tilde{\mu}_i^2}{\tilde{\sigma}_i^2} + \log \tilde{\sigma}_i^2 + \log(2\pi) \right) \right) \\ &\quad + \frac{1}{2} (\boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu} + \log|\Sigma|) \end{aligned} \tag{10}$$

In the above we have broken this equation up into three lines to clarify that the normalisation term  $\log Z$  has contribution from the prior  $p_0(\mathbf{x})$  (first line), the approximate factors (the second line), and the full approximation  $q(\mathbf{x})$  (third line). An intuitive interpretation is that the integral of  $q(\mathbf{x})$  is equal to the product of all the contributing factor masses (the prior and the approximate factors  $\tilde{t}_i$ ), divided by what is already counted by the normalisation constant of the final approximation  $q(\mathbf{x})$ . Similar reparameterisations of these calculations help to ensure numerical stability and quicker, more accurate computation. In this reparameterisation, the only cubic time operation is the computation of the log determinant and a cholesky factorization of  $K$ — we never have to invert or solve for  $\Sigma$  explicitly during the algorithm.